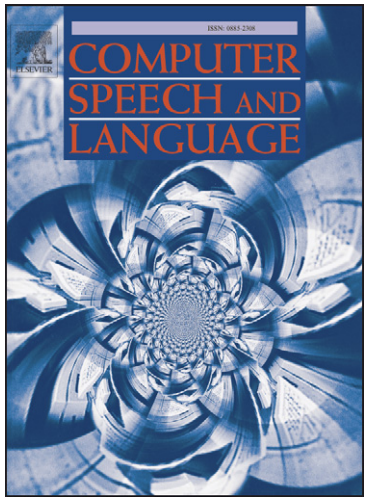

Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

PII: S0885-2308(09)00070-9
DOI: 10.1016/j.csl.2009.10.001
Reference: YCSLA 434

To appear in: *Computer Speech and Language*

Received Date: 9 July 2008
Revised Date: 14 August 2009
Accepted Date: 22 October 2009

Please cite this article as: Huet, S., Gravier, G., Sébillot, P., Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition, *Computer Speech and Language* (2009), doi: 10.1016/j.csl.2009.10.001

Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition ⋆

Stéphane Huet [1], Guillaume Gravier [*,2], Pascale Sébillot [3]

IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France

Abstract

Many automatic speech recognition (ASR) systems rely on the sole pronunciation dictionaries and language models to take into account information about language. Implicitly, morphology and syntax are to a certain extent embedded in the language models but the richness of such linguistic knowledge is not exploited. This paper studies the use of morpho-syntactic (MS) information in a post-processing stage of an ASR system, by reordering N-best lists. Each sentence hypothesis is first part-of-speech tagged. A morpho-syntactic score is computed over the tag sequence with a long-span language model and combined to the acoustic and word-level language model scores. This new sentence-level score is finally used to rescore N-best lists by reranking or consensus. Experiments on a French broadcast news task show that morpho-syntactic knowledge improves the word error rate and confidence measures. In particular, it was observed that the errors corrected are not only agreement errors and errors on short grammatical words but also other errors on lexical words where the hypothesized lemma was modified.

Key words: speech recognition, morpho-syntax, tagging, confidence measure

⋆ Some of the material presented here was previously published in conference proceedings (Huet et al., 2007) and (Huet et al., 2008).

* Corresponding author.
Email address: guillaume.gravier@irisa.fr (Guillaume Gravier).
[1] Currently at the DIRO - Université de Montréal
[2] CNRS
[3] INSA de Rennes

1 Introduction

Word-based language models (LMs) are used in all current automatic speech recognition (ASR) systems. This kind of models is a convenient way to introduce knowledge about the language thanks to statistical methods based on the frequency of occurrence of sequences of words. However, these LMs only indirectly take into account high-level linguistic information about the language, while more explicit linguistic knowledge could be included, in particular morphology and syntax. The joint use of these two knowledge sources to improve transcription is the focus of this paper.

As a preliminary study to motivate this work, we analyzed the transcription errors on a 30 minute excerpt of French broadcast news in order to quantize the potential gain using knowledge about syntax or morphology. The excerpt was transcribed using a 4-gram language model and exhibits a word error rate (WER) of 17.8 %. The first observation that can be made is that several parts of the transcript were not syntactically correct, errors on short grammatical words being particularly frequent as they concern more than one out of five sentences[4]. The most problematic words were the auxiliary verb forms "*a*", "*ont*" and "*est*", the prepositions "*dans*", "*en*", "*de*" and "*à*", the conjunctions "*et*", "*ou*" and "*que*" and the determiners "*les*", "*des*", "*ces*" and "*ses*". Secondly, agreement errors account for 11.7 % of the errors, where the recognition hypothesis is an homophone of the actual word, *e.g.* a confusion between "*estivale*" and "*estivales*". However, some of these errors are very difficult to correct, particularly those that require to resolve anaphora or to establish dependencies between distant words. The percentage of errors related to agreement is still 6.1 % when solely counting gender and number agreement errors that can be corrected by considering within sentence relations, which corresponds to 1.1 point in the WER. This analysis of the errors clearly emphasizes that morphology and syntax are valuable knowledge sources to correct typical errors of an ASR system and that the two are highly intricated.

In this work, we propose a new method to jointly include knowledge on morphology and syntax in order to improve the word error rate. Previous studies have already considered morphology. For instance, the use of sub-words (Kurimo et al., 2006) or of a set of features (Vergyri et al., 2004), instead of words as basic units for ASR, leads to a reduction of the WER for agglutinative languages, with a dramatic reduction of out-of-vocabulary words. However, the increased number of units in a segment prevents these methods from using LMs with a long span history. Information about syntax was also previously

[4] By misuse of language, apart from Section 3 where the segmentation problem is discussed, the term "sentence" is used in this paper to name a word sequence delimited by acoustic cues rather than linguistic considerations.

added to the ASR process through the use of a syntactic analyzer to condition probabilities of LMs (Chelba and Jelinek, 2000). Nevertheless, the conception of robust syntactic parsers for oral documents still remains a difficult issue. Maltese and Mancini (1992) consider both morphology and syntax relying on two language models, one on lemmas and one on parts-of-speech (POS). Interpolating word-level language models with class-based language models on morpho-syntactic classes (Maltese et al., 2001) is another example of the joint use of both linguistic resources.

We propose a new approach to include morpho-syntactic (MS) information in a post-processing stage of speech decoding, by reordering N-best lists. Our method relies on parts-of-speech, which in our case are grammatical classes (*e.g.* verb, noun, preposition...) along with morphological information [5]. The idea is first to automatically tag N-best lists, which allows the disambiguation of each hypothesized word according to its MS class. Although the resulting information is more basic than the one induced by syntactic parsers, it can be reliably inferred by automatic taggers, even on spoken documents with transcription errors, as shown experimentally in this article. A morpho-syntactic score is then computed at the sentence level and combined with the acoustic and language model scores. The combination of scores rather than models, along with the limited number of tags, makes it possible to take into account long distance dependencies with a 7-gram class-based LM. Finally, the new score including morpho-syntax is used to reorder N-best lists according to several usual decoding criteria, or to compute confidence measures based on posterior probabilities.

This paper is organized as follows. Previous works on the use of morphology or syntax in ASR are presented in Section 2, with an emphasis on the French language. In Section 3, the ability of automatic taggers to deal with spoken language and misrecognized words is demonstrated. Section 4 presents our contribution to use MS information to rerank N-best lists. The experimental setup along with implementation details are provided in Section 5 while results on a French broadcast news transcription task are given in Section 6. Finally, the use of morpho-syntax for confidence measure computation is presented in Section 7.

[5] Throughout the paper, we will not limit the notion of part-of-speech to the sole syntactic behavior but rather to syntactic and morphological behavior of lexical items. We will therefore use part-of-speech and morpho-syntax as synonyms.

2 Related works

In this section, related studies that include information about morphology or syntax are presented. Studies specific to the French language are detailed in a second part of the section after an overview of the language specifics.

2.1 Improving language models with linguistic knowledge

Parts-of-speech, in the general sense, are commonly used as morphological and syntactic knowledge that can be associated with words. A popular method to take into account POS relies on class-based N-gram models, where each class represents a POS (Jelinek, 1990; Maltese and Mancini, 1992), to compute the probability of a word sequence w_1^n from the possible corresponding classes c_1^n according to

$$P_{class}[w_1^n] = \sum_{c_1 \ldots c_n} \prod_{i=1}^{n} P[w_i|c_i] P[c_i|c_{i-N+1}^{i-1}] \ . \quad (1)$$

Class-based models are usually linearly interpolated with word-based LMs as both models are complementary in the way the language is considered (Maltese et al., 2001). However, even when combined with word-based LMs, class-based LMs do not usually result in a significant improvement in terms of perplexity or of WER (Weintraub et al., 1996).

To get over the limited success obtained with classical class-based LMs, Heeman (1999) proposes an approach based on a 3-gram LM defined over word/tag pairs rather than words, the recognition problem being to find out the best joint word and POS tag sequence. This method results in a significant word error rate reduction. However, due to the drastic increase of entries in the LM, it requires a very large amount of training data and heavily relies on smoothing techniques to make up for the lack of data.

Methods resorting to syntactic analyzers have also been developed to include linguistic information. In the case of applications where a constrained grammar can be considered, such as answering questions about a restaurant in a city, probabilistic context-free grammars have been widely used (Jurafsky et al., 1995). Other methods were applied to less constrained domains in order to take into account dependencies through syntactic analysis. For instance, Chelba and Jelinek (2000) condition the LM probabilities by the heads of the syntactic constituents previously seen, relying on the construction of syntactic trees. Another method developed by Wang et al. (2004) relies on tags with various information, including syntactic constraints between words of a sentence.

As far as morphological knowledge is concerned, its interest for ASR has increased a lot in the recent years, particularly for morphologically rich languages such as Turkish, Arabic or Hungarian. The main benefit of morphology is to deal with the high number of out-of-vocabulary words that these languages exhibit with respect to languages such as English. Morphology in ASR often relies on the definition of sub-word units employed in the speech decoding process instead of words. These sub-word units are either obtained from manual resources or, in the most recent studies, inferred by automatic methods. Morphology was successfully applied to agglutinative languages (Kurimo et al., 2006; Hirsimäki et al., 2006); it was also studied for other languages such as English (Huckvale and Fang, 2002). An alternative to resort to morphology relies on the use of sets of features derived from words, instead of words. These features (or factors) can include morphological, syntactic and semantic information (Vergyri et al., 2004).

2.2 Automatic speech recognition for the French language

Though not as morphologically rich as other languages such as Semitic or Finno-Ugrian languages, the French language is a relatively highly inflected language with masculine, feminine, singular and plural forms for adjectives and with many conjugation forms for verbs. However, contrary to other Romance languages such as Spanish or Italian, the inflectional process of a simple form very frequently leads to homophones. For instance, the past participle "*caché*" (hidden, singular masculine form) admits several homophones all derived from the same verb: the singular feminine form "*cachée*", the plural masculine form "*cachés*", the plural feminine form "*cachées*" and the forms of two other modes of the verb, the infinitive "*cacher*" and the indicative present "*cachez*". This particularity increases dramatically the number of homophones in the lexicon of the ASR system, thus complicating transcription. In (Gauvain et al., 1994), it was measured that 75 % of the words of an excerpt from a French newspaper had at least one homophone, whereas this number decreased to 23 % for an excerpt of the same size from an English newspaper.

Different methods were developed for the French language in order to deal with the inflectional process. Class-based language models were conceived by grouping together words associated with a same lemma in a class (El-Bèze and Derouault, 1990) or by automatically deriving classes so as to group together words that are statistically similar (Jardino, 1996). A significant decrease of the perplexity with respect to word-based models was reported but these studies were never tested in a broadcast news automatic transcription system. Using class-based approaches enables LMs to consider long distance dependencies in a sentence, in particular regarding agreements in gender and number. In (Lavecchia et al., 2006), a cache-based LM was used along with

classes related to the gender and the number of words. In (Zitouni et al., 2003), again with the idea of dealing with long span dependencies, linguistic variable length sequences were taken into account in the computation of the LM probabilities. These two methods resulted in an improved WER with respect to a classical word-based LM on a French read speech corpus. Finally, a last approach relies on a particular data representation made of several sentence hypotheses, all homophones with the best sentence hypothesis (Béchet et al., 1999; Gauvain et al., 2005). Scoring these homophone hypotheses with a word-based LM and another LM using POS improved the WER from 10.7 % to 10.5 % to transcribe French broadcast news (Gauvain et al., 2005).

All of these methods are mainly related to the high number of homophones in the French language and mostly rely on morphology. However, as mentioned in the introduction, a fair number of errors, due to small grammatical words, are related to syntax issues. We therefore propose a new approach, relying on morpho-syntactic classes to rerank sentence hypotheses, that combines syntax and morphology in a post-processing stage. The method exploits a sentence morpho-syntactic score derived from POS tagging, the MS score being combined with the acoustic and word-level LM scores to rerank the sentence hypotheses. Contrary to many approaches that integrate the MS knowledge in the LM, defining a separate sentence level MS score allows for the use of a long-span model able to deal with dependencies between distant words not taken into account by a word-based 4-gram LM.

3 Morpho-syntactic tagging of automatic transcriptions

In order to define a morpho-syntactic score at the sentence level, the first step consists in tagging each sentence hypothesis. Morpho-syntactic tagging, which aims at finding out a sequence of MS tags relevant for a given word sequence, is a widely used technique in natural language processing and taggers are now considered reliable enough. However, most experiments were carried out on written text while spoken corpora on the contrary have been seldom studied (Valli and Véronis, 1999). However, oral language has specifics that are likely to disturb taggers, such as repetitions, revisions or fillers, known as disfluencies, or even a possible unusual syntax. Moreover, ASR transcripts raise additional difficulties as they are segmented according to acoustic cues rather than linguistic cues. They also lack punctuation, and, in the case of some ASR systems such as our, capitalization. Transcription errors may also impact text processing techniques. Thus, before using MS information, we demonstrate that such noisy texts can be reliably tagged. This section first describes the methodology chosen to implement a POS tagger, and evaluates the relevance of the tagger for automatic transcripts.

3.1 Tagger elaboration

The decision to develop a specific tagger was made so as to be more flexible in our experiments, and to limit both the vocabulary and the tag sets. The MS tagger is based on the popular method of hidden Markov model (HMM). Although this technique is quite simple and less recent than others such as support vector machines or maximum entropy models, a study previously showed that their performances were similar (Brants, 2000). HMM-based taggers are also fast enough to tag many hypotheses. The resources (*i.e.*, the lexicon and training corpus) necessary to design the MS tagger are firstly described along with tokenization issues; next, focus is put on the different tag sets tested; the tagger itself is finally presented before describing the post-processing step applied to tag sequences.

3.1.1 Resources

To develop a HMM-based MS tagger, a lexicon which contains the associations between words and their possible tags, and a tagged training corpus to learn the probabilities of POS tag sequences are required.

Concerning the first resource, advantage was taken of the use of a fixed lexicon by the ASR system to restrain the lexicon of the tagger to the sole words of the ASR system vocabulary. This interesting property eliminates the problem of unknown words that would require more complex methods. However, some differences were introduced between the two lexicons, because of their different goals. On the one hand, the lexicon used to transcribe aims at maximizing the lexical coverage with a fixed number of entries, which led to remove capitalization or to add words that belong to locutions, rather than the locutions themselves. On the other hand, the lexicon of the tagger aims at having units that fit the analysis of the grammatical classes of a word sequence. Although automatic transcripts were not capitalized before tagging—which would require to solve ambiguities—, 20 multi-words were added to the lexicon of the tagger with respect to the ASR system one. Among them were included a few locutions such as "*parce que*" (*because*), "*c'est-à-dire*" (*that is*), or Latin locutions like "*a priori*". They were selected so as to be automatically identified without risk of errors in resolving ambiguities. Some named entities were also added, such as the names of the French radio channels frequently occurring in the corpus. The addition of these locutions requires to tokenize the transcripts before tagging.

The corpus used to train the MS tagger is a 200,000-word extract from the ESTER training corpus, which consists of French radio broadcast news. This corpus contains prepared and spontaneous speech, and is therefore more rele-

vant than written language corpora to learn a tagger for spoken documents. It has been manually transcribed, including punctuation, by human annotators. A reference tagging was established by manually correcting the output of the Cordial[6] automatic tagger, one of the best taggers available for French which was already successfully applied to spoken documents (Valli and Véronis, 1999; Gendner and Adda-Decker, 2002). The tagged corpus was processed to have a format similar to a transcript: numbers were rewritten into letters, punctuation marks, as well as capitalization, were removed, and the corpus was segmented into breath-groups according to the reference transcripts. A breath-group represents the sequence of words uttered between two breath intakes and is therefore mostly defined according to acoustic cues. Note that the breath group and sentence segmentations greatly differ: on the 200,000-word excerpt used, only 41.7 % of the ends of breath-groups correspond to sentence endings.

3.1.2 Tag sets

Rather than being limited to the main grammatical classes—adverb, adjective, noun, verb, determiner, pronoun, preposition and conjunction—, which would have resulted in more robust tagging, the tag set was chosen in order to fit the requirements of ASR for French. As mentioned in Section 2.2, confusions between genders, numbers, tenses or moods are common in French. Consequently, information about gender and number were added for nouns, adjectives and determiners; gender, number and person are associated with pronouns; number, person, mood and tense are indicated for verbs. Besides, in order to train long-span tag-based LMs, the number of selected classes was restrained with respect to very rich tag sets commonly used for French. Thus, distinction between demonstrative and possessive determiners, between ordinal and qualifying adjectives, or information about cases for pronouns were not considered as relevant to correct misrecognized words, and were not taken into account.

Three tag sets were defined. A first one, called $\mathcal{T}_{\text{orig}}$ and made of 93 tags, has separate MS tags for each of the main grammatical classes. A second one, named $\mathcal{T}_{\text{red}}$, reduces this number of tags in order to focus on morphological information about number and gender. For this reduced tag set, pronouns are split into three categories of tags: personal pronouns share the same information as verbs (person and number), relative pronouns have their own tags, and the other pronouns receive the same inflectional attributes as adjectives and nouns. For verbs, fewer conjugation categories were also considered. Finally, the tag set denoted $\mathcal{T}_{\text{ext}}$ extends $\mathcal{T}_{\text{orig}}$ by associating the 100 most

[6] Distributed by the *Synapse Développement* corporation: `http://www.synapse-fr.com/`.

Table 1
Distribution of the number of tags according to the 8 main grammatical classes for the two tag sets $\mathcal{T}_{\text{orig}}$ and $\mathcal{T}_{\text{red}}$, regardless of interjections, cardinals and symbols.

	adverb	adjective	noun	pronoun	verb	determ.	prep.	conj.
$\mathcal{T}_{\text{orig}}$	1	4	11	15	51	5	1	2
$\mathcal{T}_{\text{red}}$	1	7		4	6	5	1	2

frequent grammatical words (*e.g.* "*de*", "*la*", "*le*", "*à*"...) with specific classes. Indeed, these words are problematic for ASR systems—they are short and difficult to transcribe—and for taggers—they are very ambiguous with respect to their grammatical classes. The introduction of explicit information about such words thus appears to be relevant. $\mathcal{T}_{\text{ext}}$ also differs from $\mathcal{T}_{\text{orig}}$ by its tag set for verbs, introducing specific tags for the difficult to transcribe auxiliary verbs "*avoir*" and "*être*".

The tag sets $\mathcal{T}_{\text{orig}}$, $\mathcal{T}_{\text{red}}$ and $\mathcal{T}_{\text{ext}}$ have respectively an average number of tags per word of 1.46, 1.27 and 1.33. Table 1 provides the distribution of the two first sets according to the eight main grammatical classes, $\mathcal{T}_{\text{ext}}$ having a repartition similar to $\mathcal{T}_{\text{orig}}$. In addition to the tags mentioned in Table 1, the three sets have three separate tags for interjections, cardinals and symbols—the last tag corresponding to words such as "*arobase*" (*at sign*) used to name websites.

3.1.3 Tagger

The MS tagger is based on HMMs which use a stochastic framework to find out the most probable tag sequence t_1^n for a sentence hypothesis w_1^n according to

$$\hat{t}_1^n = \arg\max_{t_1^n} P[t_1^n|w_1^n] = \arg\max_{t_1^n} P[w_1^n|t_1^n]P[t_1^n] \tag{2}$$

where the lexicon provides the correspondence between each word and its possible tags. In order to be practically tractable, the tagging problem is usually approximated (Merialdo, 1994) by

$$\hat{t}_1^n = \arg\max_{t_1^n} \prod_{i=1}^{n} P[w_i|t_i]P[t_i|t_{i-N+1}^{i-1}] \ . \tag{3}$$

$P[w_i|t_i]$ is computed from the joint count of the pair (w_i, t_i) in the training corpus and from the number of occurrences of t_i, while $P[t_i|t_{i-N+1}^{i-1}]$ is estimated using back-off and discounting. Finding out the best order N and the best suited discounting technique so as to optimize the tag accuracy was experimentally performed on 40 minutes of broadcast news transcripts. Absolute

discounting was retained for $P[w_i|t_i]$, while the original Kneser-Ney smoothing method (Chen and Goodman, 1998) with an order $N = 3$ was selected for the computation of $P[t_i|t_{i-N+1}^{i-1}]$.

3.1.4 Post-processing

For N-best list rescoring purposes, it is interesting to post-process the output of the tagger. Indeed, reducing the number of tags in a sentence hypothesis makes it possible to take into account long dependencies with the LM built from MS information. To that end, consecutive cardinals and consecutive proper names are respectively grouped into one single cardinal tag and one single proper name tag. Besides, the tag associated with filled pauses, such as "*euh*", is removed. Hence, in section 4, l will denote the number of word/tag pairs resulting from the tagging of n words—the difference coming from the tagger specific tokenization—and m the number of tags after post-processing of the tagger output.

The different steps of the tagging process, namely tokenization, disambiguation and post-processing (or tag merging), are illustrated in Figure 2.

3.2 Evaluation

A quantitative evaluation of the morpho-syntactic tagger was performed on a one-hour show from the French broadcast news corpus ESTER consisting of 11,300 words. The quality of tagging is measured in terms of tag accuracy. To do so, each tag found by the MS tagger has thus to be compared with the corresponding tag in the reference. We first describe how the tag error rate is computed for automatic transcripts before discussing results and comparing the transcript specific taggers implemented with available taggers originally designed for texts.

For automatic transcripts, finding what would be the correct tags can be challenging, even impossible when numerous misrecognized words are present. Evaluating the quality of tagging on an automatic transcript therefore relies on an alignment at the word level with the reference transcript, the computation of the tag accuracy being limited to words correctly recognized. Figure 1 illustrates this process by showing the alignment for one breath-group.

Table 2, first line, presents results for our HMM-based tagger with the $\mathcal{T}_{\text{orig}}$ tag set on manual transcripts (column 1) and on automatic transcripts (column 2) exhibiting a WER of 22.0 %. In both cases, the tag accuracy is over 95 % which is comparable to the results usually reported on written corpora. Tagging of the ASR transcript is slightly helped by the absence of out-of-

REF		HYP	
et	COO	et	COO
de	PREP	de	PREP
déterminer	VINF	déterminer	VINF
qui	PMS	qui	*PRI*
y	ADV	–	–
peut	VINDP3S	peut	VINDP3S
être	VINF	être	VINF
dans	PREP	dans	PREP
cette	DETFS	cette	DETFS
administration	NCFS	admiration	NCFS
ou	COO	–	–
qui	PMS	qui	*PRI*
ne	ADV	ne	ADV
peut	VINDP3S	peut	VINDP3S
pas	ADV	pas	ADV
y	ADV	y	ADV
être	VINF	être	VINF

Fig. 1. Alignment of the tagged ASR system output (HYP) with the reference transcript manually tagged (REF). Misrecognized words are underlined and tagging errors are in italics.

vocabulary words. Nevertheless, the comparable performance level obtained for both transcripts establishes that MS tagging is reliable, even for texts generated by an ASR system whose recognition errors are likely to jeopardize the tagging of correctly recognized words. The robustness of tagging is mostly explained by the fact that tags are locally assigned. The good performance of the HMM-based tagger remained stable when different tag sets were used with tag accuracies of 96.4 % and 96.9 % for the tag sets $\mathcal{T}_{\text{red}}$ and $\mathcal{T}_{\text{ext}}$ respectively. In the rest of this section, tag set $\mathcal{T}_{\text{ext}}$ is used as it corresponds to a standard tag set which distinguishes the main grammatical classes and disambiguates the most common grammatical words according to their POS.

As our metric puts aside misrecognized words when evaluating tagging, a manual examination of the tagger behavior for these words was performed. Though MS tags cannot be judged as correct or not on the incorrectly tran-

Table 2
Tag accuracy (in %) for three taggers with $\mathcal{T}_{\text{orig}}$ measured on a transcript without errors and on a transcript with a measured WER of 22.0 %.

transcription	manual	automatic
HMM tagger	95.7	95.7
naive tagger	90.6	90.7
Cordial	95.0	95.2

scribed parts of the corpus, it was noticed that the tagger performed quite well when a few consecutive recognition errors appear. For instance, in the example "*mais les médecins s'avoue un peu désemparés*" (but doctors admits they are a little helpless) where "*avouent*" (admit) was erroneously transcribed as "*avoue*" (admits), the tagger correctly associates the misrecognized word with a singular form, which is relevant to further correct agreement errors. In Figure 1, it can be seen that the tagger still performs well for the incorrect word "*admiration*" transcribed instead of another noun "*administration*". The second occurrence of the pronoun "*qui*" was classified as erroneously tagged since in the reference it is tagged as an interrogative pronoun rather than as a relative pronoun. However, in the absence of the conjunction "*ou*", omitted in the transcription, the tag assigned by the tagger is correct. Globally, it appears that misrecognized words are not more frequently erroneously tagged than correct words. The most common tagging errors rather concern words difficult to tag such as "*tout*" or "*que*", or locutions and complex terms that were wrongly tokenized, errors that are frequent also in ordinary texts.

The tagger designed for these experiments was compared with two other taggers. The first tagger considered is a naive one that aims at evaluating the difficulty of tagging with respect to the available resources, *i.e.*, the tagged dictionary and the training corpus. The naive tagger associates with each word the most frequent corresponding tag in the training corpus [7]. The second tagger is Cordial. The use of the naive tagger leads to good tag accuracies of over 90 % (Table 2, line 2), which are, however, significantly lower than the ones previously obtained. These results show that the use of the HMM reduces tagging errors by more than 50 %. The comparison with Cordial [8] (Table 2, last line) shows that the tagger designed for oral transcriptions exhibits performance comparable to that of a standard tagger for the French language.

[7] As an illustration, let us note that for the $\mathcal{T}_{\text{orig}}$ tag set, 71.0 % of the words in the lexicon only have a single possible tag and therefore present no tagging difficulty.
[8] The evaluations with Cordial were performed by considering as correct the confusions between a proper name tag and a common name tag; this was required to fairly compare the taggers, since Cordial relies on capital letters—missing in the analyzed texts—to detect proper names.

4 Morpho-syntactic N-best list rescoring

As shown in the previous section, POS tagging can be reliably applied to automatic transcripts, thus making it possible to tag each sentence hypothesis in a N-best list. In this section, we describe our approach to derive a morpho-syntactic score from a sentence hypothesis based on the corresponding POS tags and to combine this score with the acoustic and word-based LM scores in order to rerank the N-best sentence hypotheses.

The choice of N-best sentence hypothesis lists, as opposed to word graphs or confusion networks, is dictated by two main motivations. The first reason is that each entry of such lists can be seen as a standard text, thus permitting disambiguation of the possible POS tags. The second reason is that, since only one possible sequence of tags is considered for each sentence, models with a longer context can be used. In word graphs, disambiguation is hardly possible and one must either consider all the possible tag sequences or expand the graph such that each node representing, for instance, a trigram $w_1w_2w_3$ is replaced by all the possible tag sequences $t_1t_2t_3$. In the first case, morpho-syntactic probabilities are averaged over all the possible tag sequences while, in the last case, the size of the word graphs increases drastically with long span models.

We first define two variants of the morpho-syntactic score before discussing the use of a score function combining acoustic, linguistic and morpho-syntactic information to process N-best sentence hypothesis lists. Results are reported in Section 6.

4.1 Combined score function

Let us denote a sentence hypothesis w_1^n, and t_1^m the most likely tag sequence as determined by morpho-syntactic tagging (including tokenization and post-processing of the tag sequence as discussed in Section 3.1.4), the probability of the tag sequence being given by

$$P[t_1^m] \approx \prod_{i=1}^{m} P[t_i|t_{i-N+1}^{i-1}] \ . \tag{4}$$

Note that the number m of tags may differ from the number n of words due to the tokenization of the word sequence and the post-processing of the tags. It is also important to note that the order of the model used to compute (4) does not need to be the same as the one of the model used in (3) for disambiguation of the tag sequence and, in practice, different orders will be used.

Most ASR systems evaluate the relevance of each sentence hypothesis w_1^n given the acoustic input y_1^t from the probabilities $P(y_1^t|w_1^n)$ and $P[w_1^n]$ respectively computed by the acoustic model and the language model. In practice, the two scores are linearly combined in the log-domain according to

$$s_{\text{orig}}(w_1^n) = \ln P(y_1^t|w_1^n) + \alpha \ln P[w_1^n] + \gamma n \ , \tag{5}$$

where α is the LM scale factor and γ a word insertion penalty. This baseline score functions is also classically used for N-best list or word graphs post-processing. To account for morpho-syntactic knowledge, we extend this global sentence score by extending the linear combination to include the morpho-syntactic score $P[t_1^m]$, *i.e.*,

$$s_1(w_1^n) = \ln P(y_1^t|w_1^n) + \alpha \ln P[w_1^n] + \beta \ln P[t_1^m] + \gamma n \tag{6}$$

where β is the POS scale factor. We propose a variant of the MS score (4) which takes into account the lexical probabilities $P[w_i|t_i]$ that are traditionally incorporated in class-based LMs [9], formally given by

$$s_2(w_1^n) = \ln P(y_1^t|w_1^n) + \alpha \ln P[w_1^n] + \beta(\ln P[t_1^m] + \sum_{i=1}^{l} \ln P[w'_i|t'_i]) + \gamma n \tag{7}$$

where ${w'}_1^l$ and ${t'}_1^l$ denote respectively the sequence of words after the tokenization step required by MS tagging and the corresponding tag sequence before the post-processing step (see Section 3.1.4 for details on the post-processing step of the tagger output).

Based on one of the two sentence level score functions, as defined in either (6) or (7), N-best lists can be rescored using various criteria. Three criteria, namely maximum *a posteriori* (MAP), minimum expected word error rate (Stolcke et al., 1997), and consensus decoding on N-best lists (Mangu et al., 2000) were studied in this paper.

4.2 Decoding criteria

In the context of N-best list rescoring, maximum *a posteriori* decoding aims at finding out the most likely sentence hypothesis given the acoustic features y_1^t, which simply consists in selecting the sentence hypothesis $w^{(i)}$ whose score

[9] See (1) in Section 2.1.

$s(w^{(i)})$ is maximum, *i.e.*,

$$\widehat{w} = \arg\max_{w^{(i)}} s(w^{(i)}) \ . \tag{8}$$

Usually, the sentence score is given by (5). To account for morpho-syntactic information, we propose to use one of the two variants (6) or (7) for $s(w^{(j)})$ instead of (5).

The maximum *a posteriori* criterion, which operates at the sentence level, aims at minimizing the sentence error rate. However, in many cases, one is rather interested in minimizing the word error rate. To do so, several alternate decoding criteria have been proposed, the most popular ones being explicit WER minimization and consensus decoding. The explicit word error rate minimization criterion (Stolcke et al., 1997) consists in finding out the sentence in a N-best list that minimizes the *a posteriori* expectation of the word error rate. Consensus decoding (Mangu et al., 2000) exploits a multiple alignment of the entries in the N-best list, represented as a confusion network. Both criteria rely on the computation of the posterior probability

$$P[w^{(i)}|y_1^t] = \frac{e^{s(w^{(i)})/z}}{\sum_j e^{s(w^{(j)})/z}} \ , \tag{9}$$

where z is a normalization factor. As for the MAP criterion, the sentence level score denoted $s()$ in the above equation can be given by either (5) or one of the two variants including morpho-syntactic knowledge.

Clearly, taking into account morpho-syntax as we propose is first and foremost beneficial at the sentence level, one of the goals of morpho-syntax being the generation of more grammatical sentence hypotheses. However, the two word error minimization criteria rely on sentence posteriors computed from sentence level scores. It is therefore interesting to measure how morpho-syntactic knowledge helps word error minimization criteria which can be detrimental to the sentence error rate. Moreover, if explicit WER minimization ends up reranking a N-best list, consensus decoding can provide new sentence hypotheses not originally in the N-best list and often less grammatical—as confirmed by the usually higher sentence error rates obtained with consensus decoding. As a consequence, the proposed combined scores with POS knowledge might impact differently the two criteria.

Experimental results are given in Section 6, after a description of the experimental setup.

5 Experimental setup

This section describes how transcription is performed and presents the corpus used. Details are also provided about the implementation of two methods from the literature that use morpho-syntactic information, namely class-based N-gram model and homophone lists reranking, to which we compare.

5.1 The ASR system

The ASR system used in all our experiments is the IRENE broadcast news transcription system (Gravier et al., 2005), jointly developed by IRISA and Telecom Paris for the ESTER broadcast news transcription evaluation campaign (Galliano et al., 2005). An overview of the system architecture is given in Figure 2 and details are given in the next paragraphs.

The pre-processing step consists in detecting regions containing speech which are then segmented into breath-groups—abusively called *sentence* in the sequel—based on fillers as provided by phonetic decoding. Let us stress the fact that sentence segmentation does not rely on any syntactic and grammatical considerations, though breath intakes and grammar are somewhat related. A partitioning into speaker turns is also performed in the pre-processing step. To avoid problems due to segmentation errors, experiments for this paper were carried out with a manual pre-processing but results with automatic pre-processing are reported at the end of Section 6.2.

The speech decoding step is carried out in three passes. A first pass with fairly simple context-independent acoustic models and a 3-gram word-based LM aims at generating large word graphs. These word graphs are then rescored with more complex context-dependent acoustic models and a 4-gram LM. Rescoring word graphs is based on a MAP decoding criterion (without the use of morpho-syntax) where the maximization is limited to the set of word sequences encoded in the word graph, thus making the use of more complex models tractable. Finally, based on the transcription from the second pass and the speaker partition obtained in the segmentation step, the acoustic models are adapted for each speaker and final word graphs are obtained by rescoring the initial word graphs with speaker-adapted acoustic models. N-best lists are extracted from the final word graphs. For speech recognition purposes, lists of 100 hypotheses were used as we observed that increasing the list size to 1,000 did not yield any improvement. However, for confidence measure computation as described in Section 7, 1,000 sentence candidates are considered.

In the post-processing step of the recognition process, each entry in the N-best lists is tagged with a 3-gram class model before rescoring. The various steps

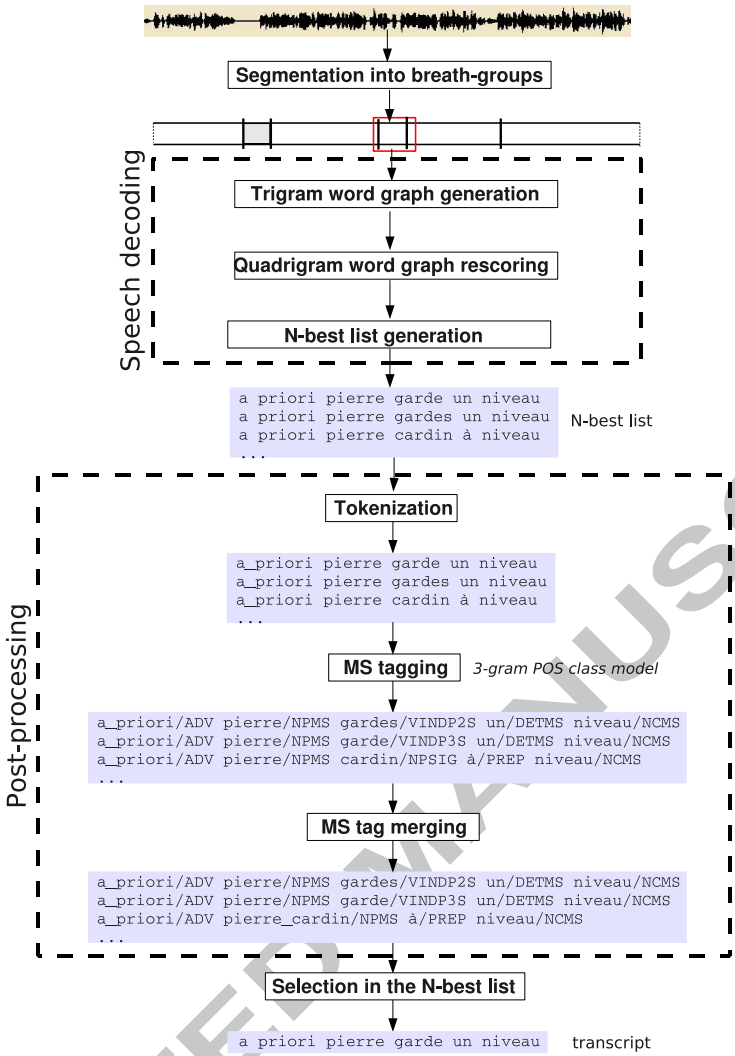

Fig. 2. Flowchart of the transcription process illustrating the 3 steps (pre-processing, speech decoding and N-best list post-processing). The post-processing step is in turn divided into the tokenization, MS tagging and tag merging steps necessary for N-best list decoding with MS information.

of the tagging process (tokenization, tagging and merging—see Section 3 for details) are depicted in Figure 2, .

5.2 Data description

As previously mentioned, experiments described in this article were carried out on the ESTER French broadcast news transcription task. A corpus of about 100 hours of manually transcribed data was used and divided into four parts (Table 3). A large part was reserved for the purpose of acoustic and language model training, 20 hours of which—*i.e.*, 200,000 words—were tagged and used

Table 3
Parts of the ESTER corpus used.

	training	development	test-a	test-b
duration	72h40	4h	4h	10h
period	1998-2003	2003	2003	2004

for the training of the N-gram tag-based models [10]. The 4-gram word-based language model, used as well for the two last decoding passes as for the post-processing of the N-best lists, was obtained by interpolating a LM estimated on 1 million words from the reference transcripts of the training set with a LM estimated from 350 million words from the French newspaper *Le Monde*.

A set of 4 hours from 4 different broadcasters makes up a development set on which parameters were optimized. Another set of 4 hours (*test-a*), recorded from the same broadcasters and during the same period as the training and the development corpora, forms a first test corpus. A last set of 10 hours (*test-b*) was used for the test; it contains 8 hours from the same 4 radio channels as the other sets and 2 hours from 2 new broadcasters. On the *test-a* set, the baseline WER is 19.7 % and the 100-best list oracle WER [11] is 11.0 %.

5.3 Class-based morpho-syntactic language model

As already mentioned in Section 2, resorting to classical class-based LMs is a common approach to take into account POS in ASR (Jelinek, 1990; Maltese and Mancini, 1992; Maltese et al., 2001). When using these models, each class corresponds to a grammatical class and the probability of a word sequence is computed according to (1). Generally, class-based LMs are used in combination with a word-based LM thanks to linear interpolation according to

$$P[w_1^n] = \prod_{i=1}^{n} [\lambda P_{word}[w_i|w_1^{i-1}] + (1-\lambda) P_{class}[w_i|w_1^{i-1}]] \quad (10)$$

where λ is the interpolation coefficient, and P_{word} and P_{class} are the probabilities computed respectively by a word-based LM—typically a N-gram LM—and a class-based LM. The conditional probabilities $P_{class}[w_i|w_1^{i-1}]$ are obtained by

[10] This corpus is actually the same as the one used to train the tagger.
[11] The N-best list oracle WER is obtained by selecting *a posteriori* the sentence hypothesis with the smallest number of errors in each N-best list.

the class-based LM thanks to

$$P_{class}[w_i|w_1^{i-1}] = \frac{P_{class}[w_1^i]}{P_{class}[w_1^{i-1}]} \tag{11}$$

where $P_{class}[w_1^i]$ is given by (1).

The interpolated language model score defined by (10) comes as a replacement for the standard word-based language in (5) for N-best list rescoring. Note that in this case, no tagging, *i.e.*, disambiguation, is performed.

5.4 Homophone N-best list generation

We emphasized in the introduction the interest of MS information to correct, among others, gender and number agreement errors for homophones. Based on this idea, an alternative data representation, lattice of homophones, has been previously suggested to take into account POS knowledge (Béchet et al., 1999; Gauvain et al., 2005). In the framework of N-best list rescoring, we implemented this method with homophone N-best lists obtained by expanding the best hypothesis as provided by the ASR system with homophones.

However, not all the possible expansions were considered. Firstly, the number of homophones for a given word can be huge. Secondly, as the homophone representation aims at correcting agreement errors, the interest of homophone expansion is restricted to the words that are likely to be concerned by such errors, *i.e.*, those that can be inflected. Consequently, the expansion of word hypotheses is limited to adjectives, nouns, verbs and personal pronouns. Besides, the list of homophones for a given word is limited to those words sharing the same lemma, *i.e.*, that differ only according to their inflection. Figure 3 illustrates the process for the sentence hypothesis "*pour construire les incontournables château*" [12] , with an agreement error on "*château*". Among the three words that can be expanded, no homophone was found in the dictionary for the verb "*construire*", the singular form of the adjective "*incontournables*" was added while the noun "*château*" was expanded with its homophone plural form "*châteaux*".

The size of homophone N-best lists ranges from a few sentences to several thousands, with an average size of 554 and a median size of 24. On the *test-a* set, the oracle WER of the homophone N-best lists is 17.7 %, much higher than the 11.0 % achieved for the standard 100-best list.

[12] The hypothesized sentence which translates into "*to build the inevitable castle*" exhibits an agreement error between "*les incontournables*" (plural form of the words "*the inevitable*") and "*chateau*" (singular form of the word castle).

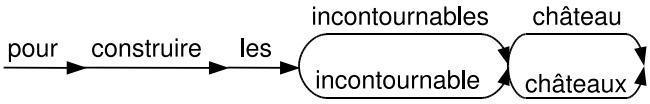

Fig. 3. Homophone expansion for the example "*pour construire les incontournables château*".

6 Transcription results

Experimental results for speech recognition are given in this section. In a first part, preliminary experiments aiming at comparing the tag sets and setting the parameters of our method are presented. The three decoding criteria are then compared before analyzing the results from a qualitative point of view. A discussion on the extension of the method to other ASR systems and languages is finally given.

6.1 Parametrization

One of the first choices that needs to be made concerns the order of the POS LM. Interestingly, it was observed in early experiments with various tag sets that transcripts are improved with a POS 4-gram model which demonstrates that the POS information is complementary to the linguistic information already embedded in the word-based LM model. The best compromise was obtained for $\mathcal{T}_{\text{ext}}$ with a 7-gram model for the computation of the morpho-syntactic score with a Kneser-Ney smoothing technique. An order of 7 is used in the remaining experiments.

In (6) or (7), the various scores are scaled with an appropriate factor. Figure 4 plots the WER on the development data as a function of the POS scale factor β for the tag set $\mathcal{T}_{\text{ext}}$, all other parameters being optimized separately for each value of β. The MAP decoding criterion is used in these experiments with the MS score (6). These results show a clear decrease of the WER for values of β between 1 and 7.

As discussed in Section 3.1.2, several tag sets can be envisaged. Results are given in Table 4 for the three tag sets where an absolute decrease of at least 0.5 of the WER was obtained with respect to the baseline system without morpho-syntax, whatever the tag set used. However, the best performance was obtained with the extended tag set $\mathcal{T}_{\text{ext}}$, which demonstrates the interest of information about grammatical classes and morphological knowledge, as well as that of the specific tags for the most frequent grammatical words. In all of the remaining experiments, the extended tag set will be used with a 7-gram model.

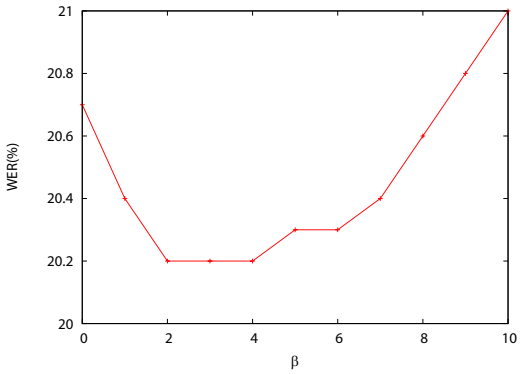

Fig. 4. WER as a function of the POS scale factor β on the development data with a MAP decoding criterion, the score $s_1(w_1^n)$ and the $\mathcal{T}_{\text{ext}}$ tag set.

Table 4
WER (%) measured on both test corpora from 100-best lists with several tag sets. The reranking of N-best lists was done with $s_1(w_1^n)$ score and with a MAP criterion.

	baseline	$\mathcal{T}_{\text{orig}}$	$\mathcal{T}_{\text{red}}$	$\mathcal{T}_{\text{ext}}$
test-a	19.7	19.1	19.2	19.0
test-b	24.7	24.2	24.2	24.0

6.2 Comparison of several approaches

We first compare several score functions for reranking N-best lists, whether standard or homophone. We then compare the three decoding criteria before discussing the significance of the results.

Results for the two score functions defined in this paper are presented in Table 5 for the MAP decoding criterion, where they are compared with a class-based LM approach and with homophone N-best list reranking. The first column reports the WER for the baseline system where no MS information is used. Results for N-best list reranking are given in columns 2, 3 and 4 respectively with $s_1(w_1^n)$, $s_2(w_1^n)$ and with a LM resulting from the interpolation of the word-level and POS-level LMs (see Section 5.3 for details). Finally, results for our method on homophone N-best list reranking, rather than standard 100-best list reranking, are presented in the last two columns for the two score functions.

The comparison of the different approaches first exhibits an absolute decrease of 0.7 for 100-best lists with $s_1(w_1^n)$ with respect to the baseline system. Using $s_2(w_1^n)$ instead leads to a more important decrease. Interestingly, we noticed that the two score functions only increase the WER for 1 out of 7 shows of *test-a* and reduce the WER for all the 18 shows of *test-b*. Besides, the improvement achieved resorting to a classical class-based LM linearly interpolated with a word-based LM is significantly less than that achieved with our method, in

Table 5
WER (%) measured on both test corpora from 100-best lists or homophone extensions, according to different scores and with MAP decoding.

	baseline	100-best lists			homophone ext.	
	$s_{\text{orig}}(w_1^n)$	$s_1(w_1^n)$	$s_2(w_1^n)$	class-based LM	$s_1(w_1^n)$	$s_2(w_1^n)$
test-a	19.7	19.0	18.8	19.1	19.4	19.4
test-b	24.7	24.0	23.9	24.2	24.4	24.4

Table 6
WER (%) measured on both test corpora with three decoding criteria.

	test-a			test-b		
	MAP	min. WE	cons.	MAP	min. WE	cons.
without MS	19.7	19.5	19.4	24.7	24.3	24.0
with MS	18.8	18.7	18.7	23.9	23.6	23.5

particular with the score function $s_2(w_1^n)$. Finally, reranking homophone N-best lists results in a limited WER gain which indicates that morpho-syntactic N-best list rescoring with our method is able to correct errors others than those on homophones. This fact will be discussed in more details at the end of the current section.

Results for N-best list rescoring with the three decoding criteria considered in this paper, namely MAP, explicit WER minimization and consensus, are given in Table 6 for the score function $s_2(w_1^n)$. The two word error minimization criteria clearly benefit from a sentence score function that includes morpho-syntactic knowledge though larger gains are achieved for the maximum *a posteriori* criterion. Consensus decoding using posteriors that include MS information yielded the best performance.

Statistical tests were carried out to measure how significant are the improvements observed, assuming independence of the errors across sentences. First, both the paired t-test and the paired Wilcoxon test, through p-values smaller than 10^{-5}, demonstrate that the scores $s_1(w_1^n)$ or $s_2(w_1^n)$ significantly improve the WER by taking into account morpho-syntax with respect to the baseline system, for the two test corpora and for all the decoding criteria. Regarding the comparison of the $s_1(w_1^n)$ and $s_2(w_1^n)$ scores with the class-based LM, results are not significantly different on the *test-a* corpus. On the *test-b* corpus, p-values lower than 5.10^{-2} and 2.10^{-3} for $s_1(w_1^n)$ and $s_2(w_1^n)$ respectively show a mild significance. Finally, improvements with respect to homophone N-best list reranking are significant for the two test sets with p-values smaller than 10^{-4}.

```
w/o MS: à part quelques MINORITÉ
w/  MS: à part quelques minorités
w/o MS: AUSSI puissant *** QUI soit
w/  MS: si    puissant qu' il  soit
```

Fig. 5. Two breath-groups transcribed without or with the use of morpho-syntax. Deletion errors are indicated by * while misrecognized words are written in capitals. The reference for the first example translates into "*apart from some minority*" w/o MS and "*apart from some minorities*" w/ MS. For the second example, the corresponding translations are respectively "*as powerful who may be*" and "*as powerful as he may be*".

6.3 Qualitative analysis of the results

In this discussion, we provide some qualitative insights on the results, regarding the robustness to speaking style and the typology of the errors corrected.

A first question raised by these experiments concerns the robustness of the method to non prepared speech, for which syntax is relaxed. Word error rates were computed separately on a short extract of 3,650 words, containing interviews with numerous disfluencies. On this extract, the baseline WER of 44.9 % is reduced to 43.7 % with $s_1(w_1^n)$ and to 43.6 % with $s_2(w_1^n)$ using the MAP criterion. This 3 % relative improvement, close to the 4 % relative improvement obtained on the entire *test-a* set, demonstrates that the method is quite robust to speaking style.

As far as the recovered errors are concerned, we observed that, as expected, many agreement errors were corrected, such as the confusion about number for the noun "*minorité*" in the first example in Figure 5. Moreover, we also noticed that hypotheses generated with the use of morpho-syntax tend to be more grammatical. The second example in Figure 5 illustrates this fact where the ungrammatical transcription for the baseline system becomes correct after changing the grammatical words "*aussi*" and "*qui*" into "*si*" and "*qu'il*". This phenomenon is assessed by the significant reduction of the sentence error rates obtained when taking into account MS information (Table 7, last line).

In order to measure quantitatively how error reductions due to MS affect words according to their grammatical class, two new metrics were defined: the lemma error rate (LER) and the lemma error rate on lexical words ($\mathrm{LER}_{\mathrm{lex}}$). The basic idea of these metrics is to ignore errors about inflections and, eventually, about grammatical words. The LER is a straightforward extension of the WER where the error rate is computed over lemmas rather than words, thus ignoring inflection errors. A comparison between the WER and the LER therefore roughly indicates the proportion of misrecognized words due to agreement

Table 7
Word error rate, lemma error rates and sentence error rate (SER) on the *test-b* corpus (MAP criterion).

	test-b		
	baseline	$s_1(w_1^n)$	$s_2(w_1^n)$
WER	24.7	24.0	23.9
LER	21.8	21.4	21.3
LER_{lex}	22.9	22.6	21.8
SER	70.5	68.8	69.1

errors. In the first example in Figure 5, the LER would be the same for the two hypotheses since "*minorité*" (minority) and "*minorités*" (minorities) share the same lemma. The LER limited to lexical words limits the computation of the LER to nouns, verbs and adjectives, therefore measuring whether morpho-syntax rather affects grammatical or lexical words. For instance, in the second example of Figure 5, the computation of LER_{lex} would be limited to the only lexical word of the hypothesis, *i.e.*, "*puissant*" (powerful).

Computing the two lemma error rates requires that the reference and automatic transcripts be tagged and lemmatized, which was done automatically using our tagger and FLEMM (Namer, 2000). Note that the automatic lemmatization step is error-prone thus introducing a bias in the LER computation. However, our tagger performs well on broadcast news, as shown in Section 3.2, and numerous tagging errors affect grammatical words, which are easy to lemmatize. For LER_{lex}, the reference and ASR transcripts are restricted to nouns, verbs and adjectives before aligning the two. Auxiliaries and modal verbs are also discarded.

Results obtained on the *test-b* corpus according to the two lemma error rates are given in lines 2 and 3 of Table 7. The comparison between the WER and the LER shows that for the baseline system, 2.9 % of the words (24.7-21.8) are correct according to their lemma, but have a wrong inflection. This figure is reduced to 2.6 % using morpho-syntax, which indicates that this knowledge corrects some agreement errors. The use of morpho-syntax leads to an absolute decrease of the WER by 0.7 or 0.8 depending on the score function used, which translates into an absolute decrease of the LER by 0.4 or 0.5 point respectively. This suggests that globally around 40 % of the gain due to morpho-syntax are related to inflection errors and 60 % correspond to changes of lemmas. These results also explain why homophone N-best lists, for which the use of morpho-syntax is limited to inflections, yield less gain than regular N-best lists.

Interestingly, a study of LER_{lex} reveals a different influence of morpho-syntactic information according to the score function used. Indeed, $s_2(w_1^n)$ leads to

a more significant decrease of $\mathrm{LER}_{\mathrm{lex}}$ than $s_1(w_1^n)$. This indicates that $s_2(w_1^n)$ tends to modify the lemmas of content words, while $s_1(w_1^n)$ acts more upon grammatical words. The greater effect of $s_2(w_1^n)$—that takes into account the intra-class probabilities $P(w_i|t_i)$—on content words is related to its inclination to select words associated with their most frequent tag.

6.4 Discussion

Several comments can be added at this point regarding the portability of the method to different ASR systems.

First of all, let us recall that a manual sentence segmentation was used in the experiments presented so far. Automatic sentence segmentation, by making segmenting based on less syntactic considerations than human might be detrimental to MS post-processing. However, the proposed approach was recently used in the ESTER 2 radio broadcasts transcription evaluation campaign [13] (Galliano et al., 2009) with a fully automatic system where a relative improvement ranging from 1 % to 3 % relative, depending on the show, was achieved. The fact that the relative improvement is slightly less than what was reported in this paper is mainly due to the fact that the vocabulary of the ASR system has changed, that of the tagger no longer matching that of the recognition system. Hence, automatic segmentation into breath-group based on fillers detected by a phone-loop does not impact morpho-syntactic rescoring as proposed in this paper.

Secondly, the ASR system used in this study though based on standard HMM-based techniques is not a state-of-the-art system for which lower WER can be expected. Even if the ASR system used in these experiments is not a state-of-the-art one, we believe that the proposed approach would also benefit to some extent to a such system. The higher error rates of our system are due to two main reasons: the acoustic model is far from the state-of-the-art (no cross-word triphones in the first pass, no telephone bandwidth model, neither CMLLR nor SAT, *etc.*) and the amount of training data for the language model is limited (less than 400M words). However, many of the errors targeted by our MS rescoring method are still present in state-of-the-art systems, though to a lesser extent. A good example is given by the errors on homophones (Gauvain et al., 2005), clearly related to the language model. Errors on short grammatical words might be reduced with better acoustic modeling but will still be present in the transcripts. On top of that, morpho-syntactic post-processing is carried out with a 7-gram class-based LM whose span is longer than that of the 4-gram

[13] The ESTER 2 evaluation campaign is a follow-up of the 2005 ESTER campaign targeting mostly radio broadcast news data with native and accented speakers, with additional non planned speech data.

word-level LMs used in state-of-the-art systems. Clearly, a long-span model should benefit to a system even if the LM has been trained on more data.

Finally, as emphasized in the previous section, the use of morpho-syntax is far from being limited to inflection errors as about 60 % of the errors that recovered result in a change of lemma. This suggests that less morphologically rich languages can benefit from MS information. Furthermore, in complementary experiments not reported in this paper, we successfully applied our method to rerank N-best lists generated by an on-line handwriting recognition system for the English language (Quiniou et al., 2005). Though not directly comparable to the speech recognition field, this last result suggests that morpho-syntactic post-processing of N-best lists should be of interest for less inflectional languages.

7 Confidence measures

The experiments reported in the previous section demonstrated, both quantitatively and qualitatively, the interest of morpho-syntax to significantly improve ASR transcripts by post-processing N-best lists. In particular, we showed that morpho-syntactic information is effective to compute the sentence posteriors that are used in decoding with WER minimization criteria. These same sentence posteriors on N-best lists are also commonly used for the computation of confidence measures (Rueber, 1997) which should therefore also benefit from morpho-syntax information.

In this section, the appropriateness of morpho-syntax to detect recognition errors is first highlighted before evaluating confidence measures that exploit sentence posteriors computed over the previously defined score functions combining acoustic, language model and morpho-syntactic knowledge.

7.1 Interest of morpho-syntax to detect errors

To assess the interest of a morpho-syntactic score for confidence measures, the local scores $P(t_i|t_{i-6}^{i-1})$ and $P(w_i|w_{i-3}^{i-1})$ of the 7-gram tag-based LM and of the 4-gram word-based LM were plotted for a few breath-groups. These plots reveal an interesting property of the tag-based LM, as illustrated in Figure 6. Indeed, in the case of the tag-based LM, the local scores are significantly lower for erroneous words than for correct ones while with the word-based LM score, low local score were observed on correct words. This difference can be explained by the more frequent use of smoothing or back-off for the word-based LM, over a vocabulary of 64,000 words, than for the POS LM over a

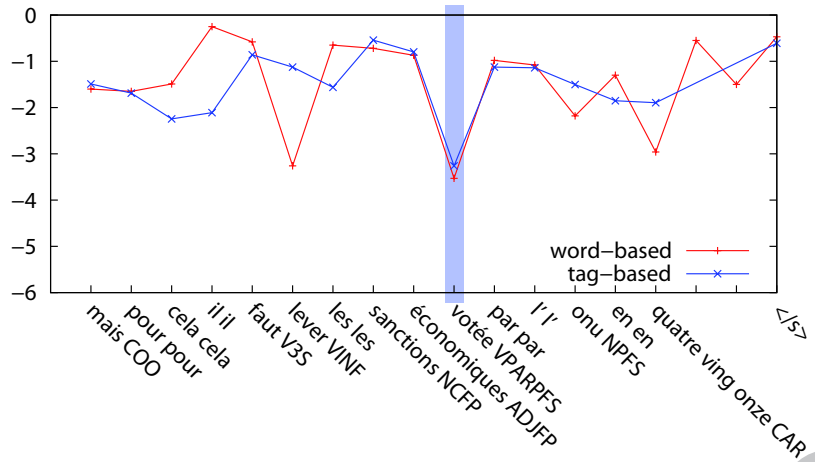

Fig. 6. Conditional log-probabilities $\ln P(t_i|t_{i-6}^{i-1})$ and $\ln P(w_i|w_{i-3}^{i-1})$ computed for each word (tag set = $\mathcal{T}_{\text{ext}}$). The shaded area over the past participle "*votée*" indicates an error.

few hundred tags.

The interest of such a property was assessed quantitatively by comparing the local scores, or a combination of the local scores, to a threshold in order to determine whether a word is correct or not. When considering solely the word-based LM, a word w_i is considered as erroneous if

$$\ln P(w_i|w_{i-3}^{i-1}) \leq \mu + a\,\sigma \quad , \tag{12}$$

where μ and σ are respectively the mean and standard deviation of the LM log-probabilities over the whole sentence. One can also consider the word-based and POS-based LM jointly and decide that a word is erroneous if

$$\ln P(w_i|w_{i-3}^{i-1}) + \ln P(t_i|t_{i-6}^{i-1}) \leq \mu' + a\,\sigma' \quad , \tag{13}$$

where μ' and σ' are defined in a similar way as for (12).

Recall and precision curves obtained by varying the parameter a are given in Figure 7 on a 30 minute excerpt from the *test-a* corpus, where recall is the proportion of correct words detected as such and precision is the proportion of correct words in the set of words whose score is above the threshold. To better study the interest of morpho-syntax for agreement errors, two tasks are considered. In task 1, all words are considered while in task 2, only words whose transcription exhibits either an agreement error or a confusion between past participle and infinitive are considered. The low precision for the second task is due to the fact that, apart for very low values of the POS log-probabilities, words found as erroneous are often detected as such for reasons not related to agreement errors or confusion between past participle and infinitive. In all

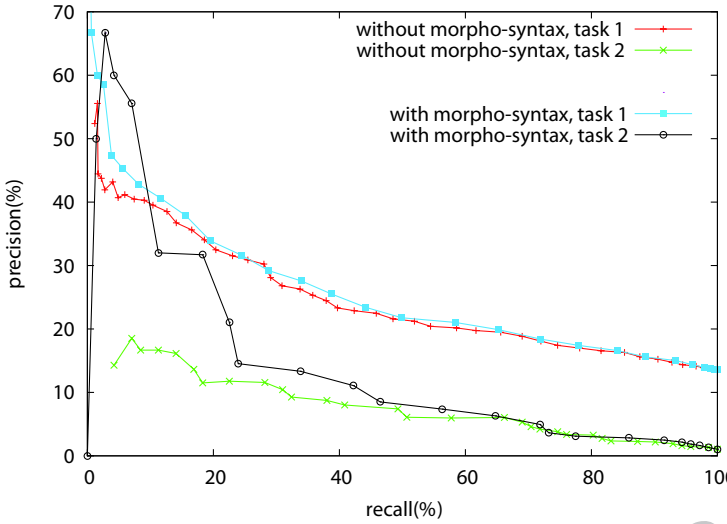

Fig. 7. Recall-precision curves for the detection of transcription errors by a word-based LM and a tag-based LM (tag set = $\mathcal{T}_{\text{ext}}$).

cases, the use of morpho-syntactic knowledge enables the detection method to improve at all recall values, especially for task 2. Nevertheless, this improvement is mainly observed for low recall values, corresponding to low values of a. This indicates that when the two LMs face a sequence not very frequent in the training corpus, the decision of the tag-based LM is more reliable than the one of the word-based LM.

7.2 Confidence measure computation

The previous results demonstrate that morpho-syntactic information can help to detect errors. We therefore study the impact of scores including morpho-syntactic information for confidence measure computation from N-best lists.

Confidence measures are computed either from the N-best lists after reranking by MAP decoding or from the confusion networks built for consensus decoding. When MAP decoding is used, the confidence measures are computed from the sentence posterior probabilities as in (Rueber, 1997). The idea is to align each alternate sentence hypothesis with the best one. The confidence measure for a word $w_i = w$ in the best sentence hypothesis is the sum of the sentence posterior probabilities over all the sentences that contain the same word w aligned with w_i. When consensus decoding is used, confidence measures are directly obtained from the highest posterior probability for each slot in the confusion network.

Normalized cross entropy values (NCE) (Siu and Gish, 1999), which measure the mutual information between correctly recognized words and the computed

Table 8
WER and normalized cross entropy measured on the two test corpora for two decoding criteria, with and without MS information.

test-a				
	MAP decoding		consensus decoding	
	w/o MS	w/ MS	w/o MS	w/ MS
WER	19.7 %	18.7 %	19.4 %	18.6 %
NCE w/o MS	0.307	0.265	0.198	-
NCE w/ MS	0.326	0.288	-	0.211

test-b				
	MAP decoding		consensus decoding	
	w/o MS	w/ MS	w/o MS	w/ MS
WER	24.7 %	23.9 %	24.0 %	23.5 %
NCE w/o MS	0.288	0.263	0.254	-
NCE w/ MS	0.305	0.275	-	0.258

confidence measures with respect to an optimal constant confidence measure, are reported in Table 8 for MAP and consensus decoding on the two test sets, using the score function $s_2(w_1^n)$. The word error rates for the various configurations tested, namely MAP/consensus decoding with/without morpho-syntax are reported on the first line of the table for each corpus. The next two lines report the NCE obtained when computing confidence measures respectively without and with morpho-syntactic knowledge. In all cases, using morpho-syntax clearly improves confidence measures, thus indicating that the MS score provides additional syntactic information not sufficiently captured by the word-based LM. It can also be noted that the NCE is lower for consensus decoding than for MAP decoding. This is due to the fact that the parameters, and in particular the scale factor z in (9), are optimized to maximize the NCE in the second case, since different parameters can be used for rescoring and for confidence measure computation in MAP decoding.

The detection error trade-off curves, such as the one plotted in Figure 8 on the development corpus with MAP decoding, confirms the interest of morpho-syntax for confidence measures. In particular, the observed improvement mainly concerns the high confidence measure values, *i.e.*, when the words most likely misrecognized are discarded. This phenomenon is in fact correlated with the improvements observed previously in Figure 7 for decisions upon recognition errors with low threshold values.

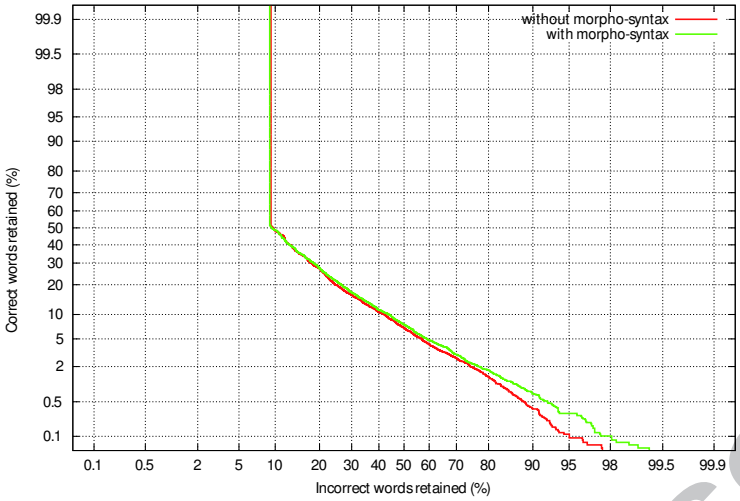

Fig. 8. Detection error trade-off curves obtained on the development corpus for a transcription using MAP reranking with morpho-syntax (WER=20.2 %, NCE without MS=0.223, NCE with MS =0.241).

8 Conclusion

We presented a study on the integration of morpho-syntactic knowledge into ASR systems and validated it for the French language. Unlike previous works that introduced morpho-syntactic knowledge at the LM level by interpolation of a class-based N-gram model with a word-based one, we combine the acoustic, LM and morpho-syntactic log scores at the sentence level. The resulting score function was found to significantly and consistently decrease the WER both with maximum *a posteriori* and explicit WER minimization decoding criteria. Besides, sentence posterior probabilities incorporating morpho-syntactic knowledge were shown to be relevant to improve confidence measures.

An analysis of the modifications induced by the use of morpho-syntactic information in transcription, as well as the observed decrease of the sentence error rate, show that this kind of information generates more grammatical transcripts. This result is particularly interesting since more grammatical transcripts should facilitate the use of natural language processing techniques for high level semantic analysis. Moreover, the study of the lemma error rate shows that, if a significant proportion of the errors corrected is related to gender and number agreements, or conjugation mistakes, morpho-syntax globally reduces the number of errors on lemmas. This last results, along with complementary experiments, suggest that the method should extend to other, less inflected, languages such as English. However, further experiments remain to be done with other languages to measure to what extent MS information can improve transcription.

Let us conclude by emphasizing the fact that the general idea of combining log-scores at the sentence level to rescore N-best lists can readily be extended to other knowledge sources such as semantics or advanced syntax. For example, a semantic score indicating whether the words in the sentence are semantically related could be considered. As far as syntax is concerned, deriving from a full syntactic parsing a score that reflects the appropriateness of the chunks in a sentence might be a way to consider syntax in an ASR system without the burden of integrating a complex algorithm such as a parser at the decoder level.

Acknowledgments

We are grateful to the two anonymous reviewers (who will recognize themselves) for their valuable comments for improving this manuscript.

References

Béchet, F., Nasr, A., Spriet, T., de Mori, R., 1999. Large span statistical language models: Application to homophone disambiguation for large vocabulary speech recognition in French. In: Proceedings of the European Conference on Speech Communication and Technology. Budapest, Hungary, pp. 1763–1766.

Brants, T., 2000. TnT - a statistical part-of-speech tagger. In: Proceedings of the Conference on Applied Natural Language Processing. Seattle, WA, USA, pp. 224–231.

Chelba, C., Jelinek, F., 2000. Structured language modeling. Computer Speech and Language 14 (4), 283–332.

Chen, S. F., Goodman, J., 1998. An empirical study of smoothing techniques for language modeling. Tech. rep., Harvard University, Cambridge, MA, USA.

El-Bèze, M., Derouault, A.-M., 1990. A morphological model for large vocabulary speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Albuquerque, NM, USA, pp. 577–580.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G., 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: Proceedings of the European Conference on Speech Communication and Technology. Lisbon, Portugal, pp. 1149–1152.

Galliano, S., Gravier, G., Chaubard, L., 2009. The ESTER 2 evaluation cam-

paign for the rich transcription of French radio broadcasts. In: Proc. 10th Conf. of the Intl. Speech Communication Association (Interspeech).

Gauvain, J.-L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., Schwenk, H., 2005. Where are we in transcribing French broadcast news? In: Proceedings of the European Conference on Speech Communication and Technology. Lisbon, Portugal, pp. 1665–1668.

Gauvain, J.-L., Lamel, L., Adda, G., Adda-Decker, M., 1994. The LIMSI continuous speech dictation system. In: Proceedings of the ARPA Workshop on Human Language Technology. Plainsboro, NJ, USA, pp. 319–324.

Gendner, V., Adda-Decker, M., 2002. Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques. In: Actes des 24èmes Journées d'études sur la Parole. Nancy, France, pp. 13–16.

Gravier, G., Yvon, F., Ben, M., 2005. IRENE, le systÃĺme irisa – enst d'indexation d'Ãľmissions radiophoniques. In: Atelier ESTER Phase II.

Heeman, P. A., 1999. POS tags and decision trees for language modeling. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, MD, USA, pp. 129–137.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language 20 (4), 515–541.

Huckvale, M., Fang, A. C., 2002. Using phonologically-constrained morphological analysis in continuous speech recognition. Computer Speech and Language 16 (2), 165–181.

Huet, S., Gravier, G., Sébillot, P., 2007. Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation. In: Proceedings of Interspeech. Antwerp, Belgium, pp. 1741–1744.

Huet, S., Gravier, G., Sébillot, P., 2008. Morphosyntactic resources for automatic speech recognition. In: Proceedings of the international conference on Language Resources and Evaluation. Marrakech, Morocco, pp. 692–698.

Jardino, M., 1996. Multilingual stochastic N-gram class language models. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA, USA, pp. 161–163.

Jelinek, F., 1990. Readings in Speech Recognition. Morgan Kaufmann Publishers, Ch. Self-Organized Language Modeling for Speech Recognition, pp. 450–506.

Jurafsky, D., Wooters, C., Segal, J., Stolcke, A., Fosler, E., Tajchman, G., Morgan, N., 1995. Using a stochastic context-free grammar as a language model for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Detroit, MI, USA, pp. 189–192.

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Saraclar, M., 2006. Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. New York

City, NY, USA, pp. 487–494.

Lavecchia, C., Smaïli, K., Haton, J.-P., 2006. How to handle gender and number agreement in statistical language models? In: Proceedings of the International Conference on Spoken Language Processing. Pittsburgh, PA, USA, pp. 1854–1857.

Maltese, G., Bravetti, P., Crépy, H., Grainger, B. J., Herzog, M., Palou, F., 2001. Combining word- and class-based language models: A comparative study in several languages using automatic and manual word-clustering techniques. In: Proceedings of the European Conference on Speech Communication and Technology. Aalborg, Denmark, pp. 21–24.

Maltese, G., Mancini, F., 1992. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, CA, USA, pp. 157–160.

Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. Computer Speech and Language 14 (4), 373–400.

Merialdo, B., 1994. Tagging English text with a probabilistic model. Computational Linguistics 20 (2), 155–171.

Namer, F., 2000. FLEMM : un analyseur flexionnel du français à base de règles. Traitement automatique des langues 41 (2), 523–547.

Quiniou, S., Anquetil, É., Carbonnel, S., 2005. Statistical language models for on-line handwritten sentence recognition. In: Proceedings of the International Conference on Document Analysis and Recognition. Seoul, South Korea, pp. 516–520.

Rueber, B., 1997. Obtaining confidence measures from sentence probabilities. In: Proceedings of the European Conference on Speech, Communication, Technology. Rhodes, Greece, pp. 739–742.

Siu, M., Gish, H., 1999. Evaluation of word confidence for speech recognition systems. Computer Speech and Language 13 (4), 299–319.

Stolcke, A., König, Y., Weintraub, M., 1997. Explicit word error minimization in N-best list rescoring. In: Proceedings of the European Conference on Speech, Communication, Technology. Rhodes, Greece, pp. 163–166.

Valli, A., Véronis, J., 1999. Étiquetage grammatical de corpus oraux : problèmes et perspectives. Revue française de linguistique appliquée 4 (2), 113–133.

Vergyri, D., Kirchhoff, K., Duh, K., Stolcke, A., 2004. Morphology-based language modeling for arabic speech recognition. In: Proceedings of the 8th International Conference on Spoken Language Processing. Jeju Island, South Korea, pp. 2245–2248.

Wang, W., Stolcke, A., Harper, M. P., 2004. The use of a linguistically motivated language model in conversational speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canada, pp. 261–264.

Weintraub, M., Aksu, Y., Dharanipragada, S., Khudanpur, S., Ney, H.,

Prange, J., Stolcke, A., Jelinek, F., Shriberg, E., 1996. LM95 project report: Fast training and portability. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Zitouni, I., Smaïli, K., Haton, J.-P., 2003. Statistical language modeling based on variable-length sequences. Computer Speech and Language 17 (1), 27–41.